BIOINFORMATICS APPLICATIONS NOTE Vol. 28 no. 23 2012, pages 3147–3149
doi:10.1093/bioinformatics/bts587

Sequence analysis

Advance Access publication October 7, 2012

ncPRO-seq: a tool for annotation and profiling of ncRNAs in sRNA-seq data

Chong-Jian Chen[1,2,3,4,5,*,†], Nicolas Servant[1,4,5,*,†], Joern Toedling[1,2,3,4,5,6], Alexis Sarazin[7], Antonin Marchais[8], Evelyne Duvernois-Berthet[7], Valérie Cognat[9], Vincent Colot[7], Olivier Voinnet[8], Edith Heard[1,2,3,‡], Constance Ciaudo[1,2,3,8,*,‡] and Emmanuel Barillot[1,4,5,‡]

[1]Institut Curie, [2]CNRS UMR3215, [3]INSERM U934, [4]INSERM U900, F-75248 Paris, France, [5]Mines ParisTech, F-77300 Fontainebleau, France, [6]Institute of Molecular Biology gGmbH, G-55128 Mainz, Germany, [7]Institut de Biologie de l'Ecole Normale Supérieure, CNRS UMR8197, and INSERM U1024, F-75248 Paris, France, [8]Department of Biology, Swiss Federal Institute of Technology, Department of Biology, S-8092 Zurich, Switzerland and [9]Institut de Biologie Moléculaire des Plantes, CNRS UPR2357, Université de Strasbourg, F-67084 Strasbourg, France

Associate Editor: Ivo Hofacker

ABSTRACT

Summary: Non-coding RNA (ncRNA) PROfiling in small RNA (sRNA)-seq (ncPRO-seq) is a stand-alone, comprehensive and flexible ncRNA analysis pipeline. It can interrogate and perform detailed profiling analysis on sRNAs derived from annotated non-coding regions in miRBase, Rfam and RepeatMasker, as well as specific regions defined by users. The ncPRO-seq pipeline performs both gene-based and family-based analyses of sRNAs. It also has a module to identify regions significantly enriched with short reads, which cannot be classified under known ncRNA families, thus enabling the discovery of previously unknown ncRNA- or small interfering RNA (siRNA)-producing regions. The ncPRO-seq pipeline supports input read sequences in fastq, fasta and color space format, as well as alignment results in BAM format, meaning that sRNA raw data from the three current major platforms (Roche-454, Illumina-Solexa and Life technologies-SOLiD) can be analyzed with this pipeline. The ncPRO-seq pipeline can be used to analyze read and alignment data, based on any sequenced genome, including mammals and plants.

Availability: Source code, annotation files, manual and online version are available at http://ncpro.curie.fr/.

Contact: bioinfo.ncproseq@curie.fr or cciaudo@ethz.ch

Supplementary information: Supplementary data are available at *Bioinformatics* online.

Received on May 24, 2012; revised on September 20, 2012; accepted on September 21, 2012

1 INTRODUCTION

Research on small non-coding RNAs (ncRNAs) has advanced tremendously over recent years as a consequence of the widespread adoption of small RNA (sRNA)-seq, which has helped to characterize members of known sRNA families, such as microRNAs (miRNAs), small interfering RNAs (siRNAs) and piwi-interacting RNA (piRNA) (Brodersen and Voinnet, 2006; Ghildiyal and Zamore, 2009). In addition, analysis of sRNA-seq data has led to the identification of several novel small ncRNA families, including heterochromatic sRNA (Rajagopalan *et al*., 2006), small nucleolar RNA (snoRNA)-derived RNAs (sdRNAs) (Taft *et al*., 2009), transfer RNA (tRNA)-derived RNA fragments (tRFs) (Cole *et al*., 2009; Pederson, 2010), transcription initiation RNAs (tiRNAs), splice-site RNAs (spliRNAs) (Taft *et al*., 2010) and enhancer RNAs (eRNAs) (Kim *et al*., 2010). Given that most of the genome is transcribed (Clark *et al*., 2011), further small ncRNA families are still probably hidden in unannoated regions, awaiting detailed exploration. Despite this, most of the existing sRNA-seq analysis tools only focus on miRNAs (Hackenberg *et al*., 2011; Ronen *et al*., 2010), whereas some other tools are applicable to the prediction of special or general siRNA loci (Hardcastle *et al*., 2012; MacLean *et al*., 2010; Stocks *et al*., 2012). As far as we are aware, only two approaches—SeqCluster (Pantano *et al*., 2011) and DARIO (Fasold *et al*., 2011)—are currently available for annotating and classifying whole sRNA-seq data in an unbiased way. SeqCluster not only carefully groups non-miRNA reads into units but also provides classification and annotation of unit-sRNAs. DARIO web server contains both sRNA annotation and prediction engines, but cannot be used to process repeat-associated sRNAs, owing to the size limitation of data submission. Furthermore, these tools perform mostly gene-based analyses, providing information about read mapping in each ncRNA gene/region. However, to systematically investigate small ncRNA species in a given annotation family (i.e. miRNA, other sRNAs families or repeats families), profiling analyses, which refers to detailed descriptions of diverse features of read distribution in annotation families, are necessary. The ncRNA PROfiling in sRNA-seq (ncPRO-seq) pipeline circumvents these limitations by providing detailed information on all types of small ncRNAs and identifying unannotated regions that are significantly enriched in matching sRNAs.

*To whom correspondence should be addressed.
†The authors wish it to be known that, in their opinion, the first two authors should be regarded as joint First Authors.
‡The authors wish it to be known that, in their opinion, the last two authors should be regarded as joint Last Authors.

2 METHODS

The workflow of the ncPRO-seq pipeline is composed of five main steps: input pre-processing, read mapping, read annotation, annotation analyses and enrichment analyses (Supplementary Fig. S1). The pipeline is able to handle sequence files generated by Solexa, SOLiD and 454 sequencing technologies and alignment files in BAM format as inputs. For sequence file inputs, the pipeline provides several ways to access the basic properties of sequencing reads, such as distribution of read length, positional base, mean positional quality score and GC content. An optional step can be introduced, for both sequence and alignment inputs, by merging reads with identical sequence into non-redundant read groups, which dramatically improves the performance of the pipeline. Bowtie (Langmead *et al*., 2009) is used to align reads to the reference genome. The mapping statistics and length distribution of mapped reads are computed to summarize mapping information. All reads, including those matching multiple genomic locations (up to a user-defined threshold), are kept for subsequent analyses and weighted by the number of mapping sites. To find overlaps between read alignments and genomic annotations, i.e. to match reads to ncRNA families according to genomic coordinates, BEDTools (Quinlan and Hall, 2010) is used. Annotation files for 15 species (including metazoans and plants), based on miRBase (Kozomara and Griffiths-Jones, 2011), Rfam (Gardner *et al*., 2011), RepeatMasker, splice site and tRNAs (Dreszer *et al*., 2012), are precomputed and now available for download in the ncPRO-seq web site. Custom annotation files in gff3 format are also acceptable, which will be considered as user-defined ncRNA families and be processed in the same way as known ncRNA families. Four types of operations are provided to help users focus on subregions or regions flanking given annotations (Supplementary Fig. S2). Analyses are then performed for reads that have 100% overlap with annotations. Reads annotated as major annotation families, as well as each family of Rfam and repeats, are counted and plotted to provide a systematic view of read annotations. For each annotation family, the table containing the expression of reads in each single member and the track file for visualization are created. To obtain detailed family-based information, read profiling and read logos for each family are computed. Read profiling is represented by the distribution of read coverage along the consensus sequence and by the distribution of read length (Supplementary Fig. S3). Read logos describe the positional base bias in all reads mapped in the family. For reads that cannot be annotated as known features, a sliding-window process coupled with model fitting is performed to identify regions that are significantly enriched with such reads (Toedling *et al*., 2010).

3 IMPLEMENTATION

The ncPRO-seq pipeline can be used in a Linux/Unix-like operating system, where several required softwares have been pre-installed. Two different ways of running the pipeline are provided, a stand-alone command line version and a local web interface. Both versions require a configuration file to set up the analysis parameters. After editing the configuration file, the user can easily run the complete analysis workflow with a simple command line. The pipeline is modular and sequential, which allows the user to focus on a specific part of the pipeline without running the complete workflow. The pipeline also provides a local web interface (Supplementary Fig. S4), which offers the possibility to set up the parameters of the configuration files in a user-friendly way, to run the pipeline. Finally, an online version for small datasets and tests is also available at http://ncpro.curie.fr/online.html.

4 OUTPUT

After running ncPRO-seq, an html report is generated, facilitating the visualization of the results that are organized under different web tabs. Besides essential tabs containing information about pipeline performance, tables and figures of read quality controls, mapping and annotations, each annotation family specified by the user has an independent tab displaying the results of detailed profiling analyses. All images and tables generated by the pipeline can be visualized at high resolution or downloaded for further analysis. An example of a report from a test library is available at http://ncpro.curie.fr/results.html.

5 CONCLUSION

The ncPRO-seq pipeline provides a comprehensive approach for the annotation and prediction of small ncRNAs in sRNA-seq data. The pipeline can analyse different annotation families rather than just focusing on miRNAs. Various sequence and alignment inputs from different metazoan and plant genomes are supported. The great advantage of the pipeline over others is the capacity to perform profiling analyses of annotation families, which is critically important to investigate known small ncRNA families and to define novel small ncRNA families. The significant regions predicted in the pipeline can be further interpreted to identify novel miRNA loci and siRNA clusters.

Funding: This research was funded by Institut Curie, INCa 'Gepig', and by grants from ANR (RNA ES, E.H., O.V.), ANR (EPIMOBILE, V.C.), ERC ('Frontiers of RNAi', no. 210890, O.V.), ERC (no. 250367, E.H.) and SYBOSS (no. 242129, E.H.). E.H. and V.C. are members of the European Network of Excellence 'EpiGenesys'. This work was supported by a post-doctoral fellowship of Federation of European Biochemical Societies (FEBS) to C.C. The funding bodies had no role in the study design, data collection and analysis, decision to publish or preparation of the manuscript.

Conflict of Interest: none declared.

REFERENCES

Brodersen,P. and Voinnet,O. (2006) The diversity of RNA silencing pathways in plants. *Trends Genet*., **22**, 268–280.

Clark,M.B. *et al.* (2011) The reality of pervasive transcription. *PLoS Biol*., **9**, e1000625; discussion e1001102.

Cole,C. *et al.* (2009) Filtering of deep sequencing data reveals the existence of abundant dicer-dependent small RNAs derived from tRNAs. *RNA*, **15**, 2147–2160.

Dreszer,T.R. *et al.* (2012) The UCSC Genome Browser database: extensions and updates 2011. *Nucleic Acids Res*., **40**, D918–D923.

Fasold,M. *et al.* (2011) DARIO: a ncRNA detection and analysis tool for next-generation sequencing experiments. *Nucleic Acids Res*., **39**, W112–W117.

Gardner,P.P. *et al.* (2011) Rfam: Wikipedia, clans and the 'decimal' release. *Nucleic Acids Res*., **39**, D141–D145.

Ghildiyal,M. and Zamore,P. (2009) Small silencing RNAs: an expanding universe. *Nat. Rev. Genet*., **10**, 94–108.

Hackenberg,M. *et al.* (2011) miRanalyzer: an update on the detection and analysis of microRNAs in high-throughput sequencing experiments. *Nucleic Acids Res*., **39**, W132–W138.

Hardcastle,T. *et al.* (2012) Identifying small interfering RNA loci from high-throughput sequencing data. *Bioinformatics*, **28**, 457–463.

Kim,T.K. *et al.* (2010) Widespread transcription at neuronal activity-regulated enhancers. *Nature*, **465**, 182–187.

Kozomara,A. and Griffiths-Jones,S. (2011) miRBase: integrating microRNA annotation and deep-sequencing data. *Nucleic Acids Res.*, **39**, D152–D157.
Langmead,B. *et al.* (2009) Ultrafast and memory-efficient alignment of short DNA sequences to the human genome. *Genome Biol.*, **10**, R25.
MacLean,D. *et al.* (2010) Finding sRNA generative locales from high-throughput sequencing data with NiBLS. *BMC Bioinformatics*, **11**, 93.
Pantano,L. *et al.* (2011) A non-biased framework for the annotation and classification of the non-miRNA small RNA transcriptome. *Bioinformatics*, **27**, 3202–3203.
Pederson,T. (2010) Regulatory RNAs derived from transfer RNA? *RNA*, **16**, 1865–1869.
Quinlan,A.R. and Hall,I.M. (2010) BEDTools: a flexible suite of utilities for comparing genomic features. *Bioinformatics*, **26**, 841–842.
Rajagopalan,R. *et al.* (2006) A diverse and evolutionarily fluid set of microRNAs in *Arabidopsis thaliana*. *Genes Dev.*, **20**, 3407–3425.
Ronen,R. *et al.* (2010) miRNAkey: a software for microRNA deep sequencing analysis. *Bioinformatics*, **26**, 2615–2616.
Stocks,M. *et al.* (2012) The UEA sRNA workbench: a suite of tools for analysing and visualizing next generation sequencing microRNA and small RNA datasets. *Bioinformatics*, **28**, 2059–2061.
Taft,R.J. *et al.* (2009) Small RNAs derived from snoRNAs. *RNA*, **15**, 1233–1240.
Taft,R.J. *et al.* (2010) Nuclear-localized tiny RNAs are associated with transcription initiation and splice sites in metazoans. *Nat. Struct. Mol. Biol.*, **17**, 1030–1034.
Toedling,J. *et al.* (2010) Girafe–an R/Bioconductor package for functional exploration of aligned next-generation sequencing reads. *Bioinformatics*, **26**, 2902–2903.